SELECTION OF CREDIBILITY REGRESSION MODELS

By

Peter Bühlmann and Hans Bühlmann

ETH Zürich, Switzerland

Abstract

We derive some decision rules to select best predictive regression models in a credibility context, that is, in a 'random effects' linear regression model with replicates. In contrast to usual model selection techniques on a collective level, our proposal allows to detect individual structures, even if they disappear in the collective.

We give exact, non-asymptotic results for the expected squared error loss for a predictor based on credibility estimation in different models. This involves correct accounting of random model parameters and the study of expected loss for shrinkage estimation. We support the theoretical properties of the new model selectors by a small simulation experiment.

Keywords and Phrases

Bias-variance trade-off, Empirical Bayes, Random effects model, Shrinkage estimation, Squared prediction loss, Subset-regression.

1 Introduction

In the open market economy of today, one of the most challenging tasks of an insurer is the design of a rating system catching all relevant factors and omitting all irrelevant ones. Mathematically, this may be modelled as the endeavour of finding those covariates which lead to the best possible predictions, for example in a regression model.

In classical statistics, with a frequentist interpretation, this problem of model selection has been widely discussed, cf. Akaike (1970, 1973), Mallows (1973), Schwarz (1978), Rissanen (1989). For an overview, see also Linart and Zucchini (1986). It may be worth to recall the well known fact, that the machinery of testing hypothesis is often *inappropriate* for searching a model with optimal predictive potential. This, because generally, the optimal predictive model-structure is not equal to the true model-structure. An intuitive (and mathematically correct) reason is that unknown parameters

ASTIN BULLETIN, Vol. 29, No. 2, 1999, pp. 245-270

have to be estimated, each of them contributing (usually in an additive way) to an increase of the variance of the estimated predictor. Moreover, selection of Bayes and Empirical Bayes models is fundamentally different from selection of models with fixed effects, because the parameters themselves contribute as random variables to statistical uncertainty. And in the case of credibility models with collateral data structure, we have to consider the fact that shrinkage estimation is used.

Unlike the more traditional use of Bayes factors, the predictive point of view in Bayes model selection has been studied among others in Gelfand and Gosh (1998). As usual, the solution depends on the specification of the prior distribution. In addition, Gelfand and Gosh (1998) take the approach to find an optimal model conditioned on the data which is often a good strategy. However, in actuarial applications one typically aims for optimality on average (minimizing the overall expected loss of the insurer) instead of conditioning on the data. The Empirical Bayes predictive point of view for optimal model selection on average with the effect of estimating hyper-parameters seems unknown. Neuhaus (1985) considers a weakly related problem about the effect of additional parameters in a credibility model. But no explicit penalty for using additional model parameters is given.

We develop here an approach for selection of general linear credibility regression models which is natural in the credibility philosophy. The set-up is as follows.

(a) The expected squared loss of a predictor at a design point is used as a predictive criterion for optimality of a model.
(b) No specification of a prior distribution for structural hyper-parameters is required. The assumptions are only in terms of first and second order moments (and a linear regression structure).
(c) The focus is on best linear prediction, but still involving shrinkage estimation.

Rather than the view to condition on the data, issue (a) is more appropriate for an insurer, as already mentioned above. Point (b) leads to a 'robustness' against misspecification of the prior distribution. It is an analogue of the Gauss-Markov conditions in classical linear models leading to best linear unbiased estimators which is the issue (c). We are giving some *exact, non-asymptotic* results for the expected squared predictive loss which in turn can be estimated from the data leading to a data-driven model-selector. Besides the theoretical justification we also consider the quality of our model-selectors in a small simulation study.

Our model selection approach helps to prevent from the dangerous anti-selection phenomenon in insurance. To explain why, we briefly mention a frequently used (but bad) strategy in practice: a model for the collective data is selected with frequentist methods for fixed effects and for such a chosen model, credibility is then introduced in a second stage. This approach is missing individual structure which averages out in the collective; see also Figure 6.1 in section 6. It is clear that such collective decision making

potentially leads to anti-selection. One has to account for individual structure: this is what our approach does and it is not mislead by the collective view.

2 THE CREDIBILITY REGRESSION MODEL

Consider a class of individual risks $r \in \{1, 2, ..., N\}$, each of them with risk parameter ϑ_r and observations $X_r = (X_{1r}, ..., X_{nr})'$. For simplicity, the individual sample size n is the same for all risks r. The risk parameter is modelled, in the Empirical Bayes sense, as a random variable. The individual correct premium for 'period' i is denoted by

$$\mu_i(\vartheta_r) = \mathbb{E}[X_{ir}|\vartheta_r], \quad i = 1, ..., n.$$

We then write

$$\begin{aligned} &X_{ir} = \mu_i(\vartheta_r) + \varepsilon_i(\vartheta_r), \quad i = 1, ..., n, \\ &\mathbb{E}[\varepsilon_i(\vartheta_r)|\vartheta_r] = 0, \ \mathbb{E}\left[\varepsilon_i^2(\vartheta_r)|\vartheta_r\right] = \sigma^2(\vartheta_r)/V_i^{(r)}, \quad i = 1, ..., n. \end{aligned} \tag{2.1}$$

The interpretation of the heteroscedastic conditional variances $\sigma^2(\vartheta_r)/V_i^{(r)}$ is given through different volumes $V_i^{(r)}$ (which are just weights in the statistical terminology) for the different 'periods' i and risks r. The individual premium is assumed to follow a linear regression structure,

$$\begin{aligned} &\mu(\vartheta_r) = D\beta(\vartheta_r), \\ &\mu(\vartheta_r) = (\mu_1(\vartheta_r), ..., \mu_n(\vartheta_r))' \ \beta(\vartheta_r) = \left(\beta_0(\vartheta_r), ..., \beta_{p-1}(\vartheta_r)\right)' \end{aligned} \tag{2.2}$$

with $n \times p$ design matrix $D(p < n)$ being the same for all $r \in \{1, ..., N\}$. The specifications (2.1) and (2.2) describe the main part of the model. Thereby we tacitly assume that the risks are drawn in an i.i.d. fashion, i.e., $\vartheta_1, ..., \vartheta_N$ i.i.d. on the structural level, implying that

$$\begin{aligned} &\beta(\vartheta_1), ..., \beta(\vartheta_N) \text{ i.i.d.}, \\ &\varepsilon(\vartheta_1), ..., \varepsilon(\vartheta_N) \text{ i.i.d., where } \varepsilon(\vartheta_r) = (\varepsilon_1(\vartheta_r), ..., \varepsilon_n(\vartheta_r))', \\ &\varepsilon_i(\vartheta_r), \varepsilon_j(\vartheta_s) \text{ independent for all } i,j = 1, ..., n \text{ and } r \neq s. \end{aligned}$$

This then induces also independence between X_{ir} and X_{js} for all $i,j = 1, ..., n$ and $r \neq s$. Since the parameter vector $\beta(\vartheta_r)$ is a random variable, we have a 'random effects' linear model with replicates over different r's. In the actuarial literature, the model is known as 'credibility regression model', cf. Hachemeister (1975).

In the sequel, we assume orthogonal design together with componentwise uncorrelated regression parameters $\beta(\vartheta_r)$. Reasons for this are given in

Bühlmann and Gisler (1997). More precisely, we assume for the structure in (2.2),

$$\sum_{i=1}^{n} D_{ij} D_{ik} V_i^{(r)} = 0 \text{ for } j \neq k \quad \text{and all} \quad r = 1, \ldots, N.$$

We assume here that

$$q_i = V_i^{(r)} / V_\bullet^{(r)} \text{ is independent of } r, \text{ where } V_\bullet^{(r)} = \sum_{i=1}^{n} V_i^{(r)}.$$

Then, orthogonality as above can always be achieved by an appropriate reparametrization. Moreover, we assume uncorrelated components of the parameter vector in (2.2),

$$\mathrm{Cov}(\beta(\vartheta_r)) = \mathrm{diag}\left(\tau_0^2, \ldots, \tau_{p-1}^2\right) \text{for all } r = 1, \ldots, N,$$

and the expectation is denoted by

$$\mathbb{E}[\beta(\vartheta_r)] = \left(b_0, \ldots, b_{p-1}\right)'.$$

Finally, we also assume conditionally uncorrelated components for the errors in (2.1),

$$\mathrm{Cov}(\varepsilon(\vartheta_r)|\vartheta_r) = \mathrm{diag}\left(\frac{\sigma^2(\vartheta_r)}{V_1^{(r)}}, \ldots, \frac{\sigma^2(\vartheta_r)}{V_n^{(r)}}\right).$$

We also denote by $\mathbb{E}\left[\sigma^2(\vartheta_r)\right] = \sigma^2$, $r = 1, \ldots, N$.

2.1 The problem of selecting the optimal model

Let us ask the question how to possibly reduce the set of regression parameters (covariates)

$$\left\{\beta_0(\vartheta_r), \beta_1(\vartheta_r), \ldots, \beta_{p-1}(\vartheta_r)\right\}$$

to an optimal subset,

$$\left\{\beta_{j_1}(\vartheta_r), \ldots, \beta_{j_m}(\vartheta_r)\right\}, \quad m \leq p,$$

or how to find an optimal subset regression model,

$$\mu_i(\vartheta_r) = \sum_{k=1}^{m} D_{ij_k} \beta_{j_k}(\vartheta_r), \quad i = 1, \ldots, n.$$

Optimality is here always with respect to the expected squared error loss, see also subsection 3.1. In the sequel, we denote by $\mathcal{U} \subseteq \{0, \ldots, p-1\}$ a

subset of the regressor indices and write the corresponding subset regression model as

$$\mu_i^{\mathcal{U}}(\vartheta_r) = \sum_{j \in \mathcal{U}} D_{ij}\beta_j(\vartheta_r).$$

We then speak of the $\mathcal{U}$-submodel. Obviously, the full model $\mathcal{U} = \{0, 1, \ldots, p-1\}$ is a possible choice, too.

It is worthwhile to remember the fundamental argument for selection of fixed effects models in frequentist statistics.

(a) Each (fixed) parameter β_{j_k} that we retain needs to be estimated, hence leading to a higher variance of the estimator for $\mu_i = \sum_{j=0}^{p-1} D_{ij}\beta_j$, $i = 1, 2, \ldots, n$.

(b) On the other hand, any relevant parameter that we miss will cause a model bias in the estimator for μ_i.

All model selection procedures in frequentist statistics rely on an optimal compromise between (a) and (b), the so-called *bias-variance trade-off*, which is estimated with the observed data. We mention here various approaches such as C_p (Mallows, 1973), FPE (Akaike, 1970), AIC (Akaike, 1973), BIC (Schwarz, 1978) and MDL (Rissanen, 1989). The tool of hypothesis testing is not tailored towards optimality (with respect to some risk function such as expected squared loss) of a model. The reason is that generally, the optimal model is not equal to the true model. For the latter, testing is appropriate, but not for the first. The shift in focus from the 'true model' (associated with testing) to the optimal approximating model (associated with risk-optimal model selection) is often very fruitful in prediction problems.

For credibility models, the uncertainty of each parameter β_{j_k} is measured differently.

- By the nature of $\beta_{j_k} = \beta_{jk}(\vartheta)$ as a *random* variable (in the Bayesian sense) with a structural prior distribution.
- By the uncertainty in the relevant hyper-parameters of the structural distribution. Following the Empirical Bayes route, estimation of hyper-parameters introduces an additional source of variance contribution.

We propose in this paper model selection rules for the homogeneous credibility estimator in regression.

3 Individual predictions under known collective structure

Unless explicitly mentioned, we assume in this section that the structural parameters in the underlying full regression model are known,

$$\sigma^2 = \mathbb{E}\big[\sigma^2(\vartheta)\big],\ \tau_j^2 = \mathrm{Var}\big(\beta_j(\vartheta)\big) \quad (j = 0, \ldots, p-1).$$

Given are the risks $r = 1, \ldots, N$ and for each risk the observations

$$X_r = (X_{1r}, \ldots, X_{nr})'.$$

For any risk r from this collective we want to find an optimal $\mathcal{U}$-submodel. As we shall see, depending on the volumes of observed data for different risks, the optimal choice of a $\mathcal{U}$-submodel can vary form risk to risk. Optimality is here, as usual in credibility theory, with respect to the expected squared loss between the predictor and the observation to be predicted.

More precisely, we wish to predict a (future) observation for risk r at a (future) design point $c = (c_0, \ldots, c_{p-1}) \in \mathbb{R}^p$,

$$X_{n+1,r}(c) = \sum_{j=0}^{p-1} c_j \beta_j(\vartheta_r) + \varepsilon_{n+1}(\vartheta_r),$$

where

$$\mathbb{E}[\varepsilon_{n+1}(\vartheta_r)|\vartheta_r] = 0, \ \mathbb{E}\left[\varepsilon_{n+1}^2(\vartheta_r)|\vartheta_r\right] = \frac{\sigma^2(\vartheta_r)}{V_{n+1}^{(r)}},$$

$$\mathrm{Cov}(\varepsilon_{n+1}(\vartheta_r), \ \varepsilon_i(\vartheta_s)|\vartheta_r, \vartheta_s) = 0 \text{ for all } i = 1, \ldots, n \text{ and } r, s = 1, \ldots, N,$$

and $V_{n+1}^{(r)}$ is the volume associated to the (future) observation $X_{n+1,r}(c)$. We can think of $X_{n+1,r}(c)$ as the next observation of claims generated at the new design point c (the new set of covariates) with volume $V_{n+1}^{(r)}$: this new observation is conditionally uncorrelated from $X_{i,r}$ $(i = 1, \ldots, n)$ and of course independent of X_{is} $(s \neq r; \ i = 1, \ldots, n)$.

We allow the set of possible regressors $\{\beta_0(\vartheta_r), \ldots, \beta_{p-1}(\vartheta_r)\}$ to be too large, i.e., the full model can be overparametrized. For example, for some index $j_0 \in \{0, \ldots, p-1\}$ it could be that $\mathbb{E}[\beta_{j_0}(\vartheta_r)] = 0$, $\mathrm{Var}(\beta_{j_0}(\vartheta_r)) = 0$. Conceptually, we could alternatively write for the new observation

$$X_{n+1,r}(c) = \sum_{j \in \mathcal{T}} c_j \beta_j(\vartheta_r) + \varepsilon_{n+1}(\vartheta_r),$$

where

$$\mathcal{T} = \left\{ j \in \{0, \ldots, p-1\}; \ \mathbb{E}\left[\beta_j^2(\vartheta_r)\right] \neq 0 \right\}$$

is the index set of the true regression parameters.

The predictor based on the submodel $\mathcal{U}$ is as follows,

$$\hat{X}_{n+1,r}^{(\mathcal{U})}(c) = \sum_{j \in \mathcal{U}} c_j \hat{\beta}_j(\vartheta_r),$$

where $\hat{\beta}_j(\vartheta_r)$ denotes the homogeneous credibility estimator,

$$\hat{\beta}_j(\vartheta_r) = Z_j^{(r)} b_{jr}^X + (1 - Z_j^{(r)}) \frac{\sum_{\ell=1}^N b_{j\ell}^X Z_j^{(\ell)}}{\sum_{\ell=1}^N Z_j^{(\ell)}}.$$

The elements of this credibility estimator $\hat{\beta}_j(\vartheta_r)$ are, cf. Bühlmann and Gisler (1997),

$$b_{jr}^X = \frac{\sum_{i=1}^n X_{ir} D_{ij} V_i^{(r)}}{\sum_{i=1}^n D_{ij}^2 V_i^{(r)}} = \frac{\sum_{i=1}^n X_{ir} D_{ij} V_i^{(r)}}{S_{jj} V_\bullet^{(r)}}, \quad S_{jj} = \sum_{i=1}^n D_{ij}^2 q_i, \quad q_i = \frac{V_i^{(r)}}{V_\bullet^{(r)}},$$

$$Z_j^{(r)} = \frac{S_{jj} V_\bullet^{(r)}}{S_{jj} V_\bullet^{(r)} + \sigma^2/\tau_j^2}.$$

Here and in the sequel, a dot denotes summation over all corresponding indices.

3.1 Expected loss and model selection

The accuracy of the predictor $\hat{X}_{n+1,r}^{(\mathcal{U})}(c)$ from the $\mathcal{U}$-submodel is measured with the expected squared loss,

$$L_r^{(\mathcal{U})}(c) = \mathbb{E}\left[\left(X_{n+1,r}(c) - \hat{X}_{n+1,r}^{(\mathcal{U})}(c)\right)^2\right].$$

It depends on the risk r, the submodel $\mathcal{U}$, the future design point c and of course also on the true underlying probability distribution which is implicitly used in the expectation operator $\mathbb{E}$. It is instructive to decompose this expected loss as

$$\begin{aligned} L_r^{(\mathcal{U})}(c) &= \sigma^2/V_{n+1}^{(r)} + \mathbb{E}\left[\left(\sum_{j\in\mathcal{U}\cap\mathcal{T}} c_j\left(\beta_j(\vartheta_r) - \hat{\beta}_j(\vartheta_r)\right)\right)^2\right] \\ &\quad + \mathbb{E}\left[\left(\sum_{j\in\mathcal{T}\setminus\mathcal{U}} c_j\beta_j(\vartheta_r)\right)^2\right] + \mathbb{E}\left[\left(\sum_{j\in\mathcal{U}\setminus\mathcal{T}} c_j\hat{\beta}_j(\vartheta_r)\right)^2\right] \\ &= I + II + III + IV. \end{aligned} \tag{3.1}$$

A derivation is given in section 8. The interpretation is as follows.

I: Uncertainty of claim around the correct individual premium.

II: Uncertainty of parameters chosen in $\mathcal{U}$ which are true relevant parameters.

III: Model bias due to underparametrization.

IV: Error due to overparametrization.

Formula (3.1) is conceptual and not useful for estimating $L_r^{(\mathcal{U})}(c)$ from the data. In particular, the set of true parameters $\mathcal{T}$ is not known. By the fact

that $\beta_j(\vartheta_r) \equiv 0$ for $j \notin \mathcal{T}$, it is straightforward to write the expected squared loss as

$$L_r^{(\mathcal{U})}(c) = \sigma^2 / V_{n+1}^{(r)} + \mathbb{E}\left[\left(\sum_{j\in\mathcal{U}} c_j\left(\beta_j(\vartheta_r) - \hat{\beta}_j(\vartheta_r)\right)\right)^2\right] + \mathbb{E}\left[\left(\sum_{j\in\mathcal{U}^C} c_j\beta_j(\vartheta_r)\right)^2\right], \quad (3.2)$$

where $\mathcal{U}^C = \{0, \ldots, p-1\}\backslash\mathcal{U}$ is the complement of the set $\mathcal{U}$ with respect to the full basis model $\{0, \ldots, p-1\}$. Formula (3.2) can be explicitly rewritten in terms of the structural parameters, see section 8,

$$\begin{aligned} L_r^{(\mathcal{U})}(c) = \sigma^2 / V_{n+1}^{(r)} &+ \sum_{j\in\mathcal{U}} c_j^2\left(\frac{\sigma^2}{S_{jj}V_\bullet^{(r)}}\left[Z_j^{(r)} + \left(1 - Z_j^{(r)}\right)\frac{Z_j^{(r)}}{Z_j^{(\bullet)}}\right]\right) \\ &+ \left(\sum_{j\in\mathcal{U}^C} c_j b_j\right)^2 + \sum_{j\in\mathcal{U}^C} c_j^2\tau_j^2. \end{aligned} \quad (3.3)$$

Each parameter chosen in $\mathcal{U}$ contributes to a *variance term* $\sum_{j\in\mathcal{U}} c_j^2\left(\frac{\sigma^2}{S_{jj}V_\bullet^{(r)}}\left[Z_j^{(r)} + \left(1 - Z_j^{(r)}\right)\frac{Z_j^{(r)}}{Z_j^{(\bullet)}}\right]\right)$ *penalizing* large models, whereas the parameters in $\mathcal{U}^C$, which are not chosen, generate a (model) *bias term* $\mathbb{E}\left[\left(\sum_{j\in\mathcal{U}^C} c_j\beta_j(\vartheta_r)\right)^2\right] = \left(\sum_{j\in\mathcal{U}^C} c_j b_j\right)^2 + \sum_{j\in\mathcal{U}^C} c_j^2\tau_j^2$.

Based on (3.3), the optimal submodel is then given by

$$\mathcal{U}_{opt}^{(r)}(c) = \operatorname{argmin}_{\mathcal{U}\subseteq\{0,\ldots,p-1\}} L_r^{(\mathcal{U})}(c).$$

Remark A. In the special case with $N = 1$ (no collateral data) and $V_i \equiv 1$ $(i = 1, \ldots, n+1)$, we obtain from (3.3) the expected squared loss in classical frequentist linear fixed effects regression

$$\sigma^2 + \sigma^2 \sum_{j\in\mathcal{U}} \frac{c_j^2}{\sum_{i=1}^n D_{ij}^2} + \left(\sum_{j\in\mathcal{U}^C} c_j b_j\right)^2.$$

This formula is different from that obtained by the classical discussion aiming for models minimizing

$$n^{-1}\sum_{i=1}^n \mathbb{E}\left[\left(X_{n+1}(d_i) - \hat{X}_{n+1}^{(\mathcal{U})}(d_i)\right)^2\right].$$

This is a mean squared error averaged over the observed design points $d_i = \left(D_{i0}, \ldots, D_{i,p-1}\right)'$, cf. Weisberg (1985, App. 8A.1) in connection with Mallows C_p. In contrast, we consider here the expected squared loss at a

particular (future) design point c measuring optimal prediction at this point c. We feel that this comes closer to the aim in actuarial practice. However, unlike to selection of fixed effects models in frequentist statistics which are all of the form of a penalized *residual* sum of squares, the approach of optimal prediction at a point needs explicit consideration of a model bias term.

So far, we have assumed that the structural parameters are all known. If this is not the case, the standard approach in credibility procedures is to replace unknown structural parameters by their estimated versions. Here, the expected loss $L_r^{(\mathcal{U})}(c)$ can then be estimated from the data with the estimated structural parameters $\hat{\sigma}^2, \hat{b}_j, \hat{\tau}_j^2$ $(j = 0, \ldots, p-1)$ and $\hat{Z}_j^{(r)}$, given in (5.1)-(5.4). Formally, the estimated squared loss $\hat{L}_r^{(\mathcal{U})}(c)$ is given by the following plug-in scheme,

$$\begin{aligned} L_r^{(\mathcal{U})}(c) &= G\Big(\sigma^2, b_0, \ldots, b_{p-1}, \tau_0^2, \ldots, \tau_{p-1}^2, Z_0, \ldots, Z_{p-1}\Big), \\ \hat{L}_r^{(\mathcal{U})}(c) &= G\Big(\hat{\sigma}^2, \hat{b}_0, \ldots, \hat{b}_{p-1}, \hat{\tau}_0^2, \ldots, \hat{\tau}_{p-1}^2, \hat{Z}_0, \ldots, \hat{Z}_{p-1}\Big), \end{aligned} \tag{3.4}$$

where $G(.)$ is the function as described by formula (3.3). We prefer the notationally appealing plug-in formalism including Z_j's as arguments, although the (optimal) Z_j's are functions of σ^2, τ_j^2 and hence not intrinsic structural parameters. We select the optimal model from the data as

$$\hat{\mathcal{U}}_{opt}^{(r)}(c) = \operatorname{argmin}_{\mathcal{U} \subseteq \{0, \ldots, p-1\}} \hat{L}_r^{(\mathcal{U})}(c). \tag{3.5}$$

Like the truly optimal submodel $\mathcal{U}_{opt}^{(r)}(c)$, the estimated optimal submodel $\hat{\mathcal{U}}_{opt}^{(r)}(c)$ depends on the future design point c (where prediction is made) and on the risk r. The estimator $\hat{L}_r^{(\mathcal{U})}(c)$ is consistent as the number N of risks grows to infinity, and thus also our selection procedure.

The model selector in (3.5) is useful and quite easy to implement. However, the standard argumentation in credibility, namely to treat in a first stage the structural parameters as fixed and replace them in a second stage by their estimated versions, discards uncertainty about structural parameter estimation. The expected loss $L_r^{(\mathcal{U})}(c)$ does not account for this uncertainty and can in this sense be misleading. For practical purposes, as long as N is 'sufficiently large', the selector in (3.5) is appropriate for discriminating among 'sufficiently different' prediction models. A more detailed discussion about this issue is given in sections 6 and 7. We describe in the next section a more complicated scheme which accounts for estimation of structural parameters.

4 Unknown hyper-parameters: non-optimal and estimated credibility weights

If the structural parameters of the collective are not known, we need to estimate $Z_j^{(r)}$ $(j = 0, \ldots, p-1)$ for constructing the credibility estimator. Such an estimate is given in (5.4).

The expected loss $L_r^{(\mathcal{U})}(c)$ in subsection 3.1 is correct if the structural parameters and hence the credibility weights are known; in particular, this means that the credibility weights are optimal for unbiased linear estimation. But of course, estimated credibility weights are never exactly optimal; hence, the expected loss $L_r^{(\mathcal{U})}(c)$ is not correct. We first address the problem of obtaining the true expected squared loss for *fixed*, generally non-optimal credibility weights between 0 and 1, which are denoted in the sequel by $\tilde{Z}_j$ $(j = 0, \ldots, p-1)$. In statistical terminology, we study the expected loss for a *shrinkage* estimator with fixed shrinkage factors. The problem of treating the credibility weights as random, which is the case when they have been estimated, is more delicate and we discuss it in subsection 4.1.

The quantity in formula (3.2) that depends on the credibility weights is the variance term $\sum_{j\in\mathcal{U}} c_j^2 \mathbb{E}\left[\left(\beta_j(\vartheta_r) - \hat{\beta}_j(\vartheta_r)\right)^2\right]$. By using arbitrary, fixed credibility weights, the expected squared loss $\mathbb{E}\left[\left(\hat{X}_{n+1,r}^{(\mathcal{U})}(c) - X_{n+1,r}(c)\right)^2\right]$ is

$$
\begin{aligned}
&M_r^{(\mathcal{U})}(c) = \sigma^2 / V_{n+1}^{(r)} \\
&+ \sum_{j\in\mathcal{U}} c_j^2 \left(\sigma_{j,r}^2 \left[\tilde{Z}_j^{(r)} + (1-\tilde{Z}_j^{(r)}) \frac{\tilde{Z}_j^{(r)}}{\tilde{Z}_j^{(\bullet)}} \right]^2 + (1-\tilde{Z}_j^{(r)})^2 \frac{\sum_{\ell\neq r} \left(\tilde{Z}_j^{(\ell)}\right)^2 \left(\sigma_{j,\ell}^2 + \tau_j^2\right) + \left(\sum_{\ell\neq r} \tilde{Z}_j^{(\ell)}\right)^2 \tau_j^2}{\left(\tilde{Z}_j^{(\bullet)}\right)^2} \right) \\
&+ \left(\sum_{j\in\mathcal{U}^C} c_j b_j \right)^2 + \sum_{j\in\mathcal{U}^C} c_j^2 \tau_j^2, \\
&\sigma_{j,\ell}^2 = \sigma^2 / \left(S_{jj} V_{\bullet}^{(\ell)}\right) \quad (\ell = 1, \ldots, N). \qquad (4.1)
\end{aligned}
$$

A derivation is given in section 8. The notation $M_r^{(\mathcal{U})}(c)$ distinguishes this expected squared loss with arbitrary, fixed credibility weights from $L_r^{(\mathcal{U})}(c)$ for the optimal credibility weights. Of course, $L_r^{(\mathcal{U})}(c)$ is just a special case of $M_r^{(\mathcal{U})}(c)$ when optimizing over the credibility weights $\tilde{Z}_j$ in (4.1).

Remark B. The quantity $M_r^{(\mathcal{U})}(c)$ in (4.1) is the *exact* expected loss for the homogeneous credibility estimator with fixed arbitrary credibility weights. This is also of special interest in statistical theory: our result describes model risks in connection with *shrinkage* estimation for fixed shrinkage weights. In actuarial practice we advise to use (4.1) instead of (3.3) whenever the credibility weights are determined by other reasons and are not estimated from the collateral data structure. The difference between the expected losses in (3.3) and (4.1) can also be used in a sensitivity analysis when considering the stability of an optimal model under variation of the credibility weights around their optimal values.

Estimation of $M_r^{(\mathcal{U})}(c)$ with fixed, given $\tilde{Z}_j^{(r)}$'s can be done with plugging in the estimate $\hat{\sigma}^2$, $\hat{b}_j$ and $\hat{\tau}_j^2$ from (5.1)-(5.3).

Remark C. Another estimate of $M_r^{(\mathcal{U})}(c)$ than the one discussed above could be as follows. Plug in all estimated structural quantities $\hat{\sigma}^2$, $\hat{b}_j$, $\hat{\tau}_j^2$ and $\tilde{Z}_j^{(r)} = \hat{Z}_j^{(r)}$ from (5.1)-(5.4). But this would coincide with $\hat{L}_r^{(\mathcal{U})}(c)$ from (3.4), since the credibility weights $\hat{Z}_j^{(r)}$ are of the optimal structural form. Accounting for the randomness when plugging in such *estimated* credibility weights is given in the next subsection 4.1.

In the special case where all the volumes are the same, i.e., $V_i^{(r)} \equiv V$ $(i = 1, \ldots, n+1,\ r = 1, \ldots, N)$, the fixed credibility weights should not depend on the risk r. We denote them by $\tilde{Z}_j$ $(j = 0, \ldots, p-1)$. Then,

$$\begin{aligned} M_r^{(\mathcal{U})}(c) \equiv M^{(\mathcal{U})}(c) &= \sigma^2/V \\ &+ \sum_{j\in\mathcal{U}} c_j^2 \left(\sigma_j^2 \left(\tilde{Z}_j^2 \frac{N-1}{N} + \frac{1}{N} \right) + (1-\tilde{Z}_j)^2 \frac{N-1}{N} \tau_j^2 \right) \\ &+ \left(\sum_{j\in\mathcal{U}^C} c_j b_j \right)^2 + \sum_{j\in\mathcal{U}^C} c_j^2 \tau_j^2, \\ \sigma_j^2 &= \sigma^2/(S_{jj} n V). \end{aligned} \tag{4.2}$$

4.1 Estimated credibility weights

For simplicity we consider here the case with equal volumes as in (4.2). But all what follows can be written down straightforwardly for the general case with different volumes. When estimating credibility weights we consider,

$$\hat{Z}_j = \frac{S_{jj} n V}{S_{jj} n V + \hat{\sigma}^2/\hat{\tau}_j^2} \quad (j = 0, \ldots, p-1).$$

These estimators are consistent for the true optimal weights $Z_j = \frac{S_{jj} n V}{S_{jj} n V + \sigma^2/\tau_j^2}$ as $N \to \infty$. Direct plug in of estimated structural quantities is discussed in Remark C. We write

$$\hat{Z}_j = Z_j + \Delta_j.$$

When using $\tilde{Z}_j = \hat{Z}_j$ in formula (4.2) (ignoring first that the $\hat{Z}_j$'s are random), we obtain

$$\begin{aligned} M_r^{(\mathcal{U})}(c) \equiv M^{(\mathcal{U})}(c) &= \sigma^2/V \\ &+ \sum_{j\in\mathcal{U}} c_j^2 \left(\sigma_j^2 \left[Z_j + \frac{1-Z_j}{N} + \frac{N-1}{N}\Delta_j^2 \right] + \frac{N-1}{N}\Delta_j^2\tau_j^2 \right) \\ &+ \left(\sum_{j\in\mathcal{U}^C} c_j b_j \right)^2 + \sum_{j\in\mathcal{U}^C} c_j^2\tau_j^2, \\ \sigma_j^2 &= \sigma^2/(S_{jj}nV). \end{aligned} \tag{4.3}$$

Note the correspondence to formula (3.3) as well, since the Z_j's are the optimal fixed credibility weights. To evaluate (4.3), we need a 'reasonable' value of Δ_j^2. Approximately, $\mathbb{E}[\Delta_j] \approx 0$ and thus, the expected value of Δ_j^2 is approximately

$$\mathbb{E}\left[\Delta_j^2\right] \approx \mathrm{Var}(\Delta_j)$$

the variability of $\hat{Z}_j$. As a reasonable but non-exact value we find

$$\mathrm{Var}(\Delta_j) \approx \begin{cases} 2Z_j^2(1-Z_j)^2 \left(\frac{1}{N(n-p)} \left(1 + \frac{2\sigma^2}{\tau_j^2 S_{jj} nV} + \frac{\sigma^4}{\tau_j^4 (S_{jj}nV)^2} \right) + \frac{1}{(N-1)\tau_j^4} \left(\frac{\sigma^2}{S_{jj}nV} + \tau_j^2 \right)^2 \right), & \text{if } \tau_j^2 > 0, \\ 2\left(\frac{1}{N(n-p)} + \frac{1}{N-1} \right), & \text{if } \tau_j^2 = 0. \end{cases} \tag{4.4}$$

See section 8. Replacing the quantities Δ_j^2 in (4.3) with the values in (4.4) yields another expected loss

$$R_r^{(\mathcal{U})}(c) \equiv R^{(\mathcal{U})}(c) \text{ given by (4.3) and (4.4).} \tag{4.5}$$

Notationally, we distinguish this expected squared loss $R^{(\mathcal{U})}(c)$ for the case with estimated, random credibility weights from the one in (4.2) with fixed, arbitrary credibility weights. Note that in (4.5), the credibility weights Z_j are optimal as in (3.3) (but the expected loss is, unlike as in (3.3), for the predictor $\hat{X}_{n+1,r}^{(\mathcal{U})}(c) \equiv \hat{X}_{n+1}^{(\mathcal{U})}(c)$ with *estimated, unknown* credibility weights).

Remark D. The expected loss in (4.5) does only *partially* reflect the randomness of the estimated $\hat{Z}_j$'s for the predictor $\hat{X}_{n+1,r}^{(\mathcal{U})}(c) \equiv \hat{X}_{n+1}^{(\mathcal{U})}(c)$. Formula (4.3) treats the Δ_j's as *fixed* and we then consider afterwards statistical variability of these quantities. This route possibly misses some complicated correlation between $\hat{Z}_j$ and the individual least squares estimates b_{kr}^X ($j, k = 0, \ldots, p-1, r = 1, \ldots, N$). A non-asymptotic, exact

calculation of the expected loss when (correctly) viewing the $\hat{Z}_j$ as random variables being functions of all the observations seems very difficult.

The expected squared loss in (4.5) can be estimated by plugging in the estimated structural parameters $\hat{\sigma}^2$, $\hat{b}_j$, $\hat{\tau}_j^2$ and $\hat{Z}_j$ $(j = 0, \ldots, p-1)$ given in (5.1)-(5.4). We denote it by $\hat{R}^{(\mathcal{U})}(c)$. By the message in Remark D, we view (4.5) and its estimated version $\hat{R}^{(\mathcal{U})}(c)$ as a guide to account for effects due to estimation of credibility weights. Our simulation study in section 6 indicates that model selection based on $\hat{R}^{(\mathcal{U})}(c)$ works better than with $\hat{L}^{(\mathcal{U})}(c)$ from (3.4); in particular for discriminating among very similar prediction models. For a more detailed discussion, see sections 6 and 7.

5 Estimation of structural hyper-parameters

When estimating the expected losses in (3.3), (4.1), (4.2) and (4.5) it remains to estimate the hyper-parameters σ^2, $b_0, \ldots, b_{p-1}$ and $\tau_0^2, \ldots, \tau_{p-1}^2$.

The variance of the errors $\sigma^2 = \mathrm{Var}(\varepsilon(\vartheta_r)) = \mathbb{E}[\mathrm{Var}(\varepsilon(\vartheta_r)|\vartheta_r)]$ can be estimated with a residual sum of squares from the full basis model involving all covariates. Let

$$\mathrm{RSS}_r = \sum_{i=1}^{n} \left(X_{ir} - \sum_{j=0}^{p-1} b_{jr}^X D_{ij} \right)^2 \frac{V_i^{(r)}}{V_\bullet^{(r)}},$$

$$\mathrm{RSS} = \sum_{r=1}^{N} \mathrm{RSS}_r \frac{V_\bullet^{(r)}}{V_\bullet^{(\bullet)}}.$$

An unbiased estimator is then given by

$$\hat{\sigma}^2 = \mathrm{RSS} \frac{V_\bullet^{(\bullet)}}{N(n-p)}. \tag{5.1}$$

See section 8.

For the collective means we take the obvious estimator

$$\hat{b}_j = \sum_{r=1}^{N} \frac{V_\bullet^{(r)}}{V_\bullet^{(\bullet)}} b_{jr}^X \quad (j = 0, \ldots, p-1). \tag{5.2}$$

For the variance components $\tau_j^2 = \mathrm{Var}\big(\beta_j(\vartheta_r)\big)$, consider the sum of squares

$$W_j = \sum_{r=1}^{N} \left(b_{jr}^X - \sum_{\ell=1}^{N} b_{j\ell}^X \frac{V_\bullet^{(\ell)}}{V_\bullet^{(\bullet)}} \right)^2 \frac{V_\bullet^{(r)}}{V_\bullet^{(\bullet)}}.$$

As shown in section 8,

$$\mathbb{E}[W_j] = \tau_j^2 A + (N-1)\frac{\sigma^2}{S_{jj} V_\bullet^{(\bullet)}},\ A = 1 - \sum_{\ell=1}^{N} \left(V_\bullet^{(\ell)} / V_\bullet^{(\bullet)}\right)^2.$$

We then use as an estimator

$$\hat{\tau}_j^2 = \frac{1}{A}\left(W_j - \frac{(N-1)\mathrm{RSS}}{N(n-p)S_{jj}}\right)^+,\ A = 1 - \sum_{\ell=1}^{N}\left(\frac{V_\bullet^{(\ell)}}{V_\bullet^{(\bullet)}}\right)^2 \ (j = 0, \ldots, p-1), \quad (5.3)$$

where $u^+ = \max(u, 0)$ and RSS as above.

We then estimate credibility weights, of optimal form, as

$$\hat{Z}_j^{(r)} = \frac{S_{jj} V_\bullet^{(r)}}{S_{jj} V_\bullet^{(r)} + \hat{\sigma}^2/\hat{\tau}_j^2} \ (j = 0, \ldots, p-1), \quad (5.4)$$

with $\hat{\sigma}^2$ and $\hat{\tau}_j^2$ from (5.1) and (5.3), respectively.

As usual in model selection, it is often of little concern to ask about efficiency for estimating unknown expected losses. We just give a few comments. Due to the orthogonal design, the accuracy of the individual estimates b_{jr}^X, used in W_j, is always the same, regardless how large the full basis model is. On the other hand, the efficiency of the estimator $\hat{\sigma}^2$ does depend on the dimensionality of the full basis model. With increasing degree of overparametrizing the basis model, the estimator $\hat{\sigma}^2$ gets more inefficient. But usually, such effects are very small.

6 Simulation

We consider two related situations. In both cases, the individual sample size is $n = 10$ and the volumes are $V_i^{(r)} \equiv 1$ $(i = 1, \ldots, n+1;\ r = 1, \ldots, N)$.

The design matrix is

$$D = \begin{pmatrix} 1 & -4.9543369 & 5.2223297 & -4.534252 & 3.3658092 \\ 1 & -3.8533732 & 1.7407766 & 1.511417 & -4.1137668 \\ 1 & -2.7524094 & -0.8703883 & 3.778543 & -3.1788198 \\ 1 & -1.6514456 & -2.6111648 & 3.346710 & 0.5609682 \\ 1 & -0.5504819 & -3.4815531 & 1.295501 & 3.3658092 \\ 1 & 0.5504819 & -3.4815531 & -1.295501 & 3.3658092 \\ 1 & 1.6514456 & -2.6111648 & -3.346710 & 0.5609682 \\ 1 & 2.7524094 & -0.8703883 & -3.778543 & -3.1788198 \\ 1 & 3.8533732 & 1.7407766 & -1.511417 & -4.1137668 \\ 1 & 4.9543369 & 5.2223297 & 4.534252 & 3.3658092 \end{pmatrix}. \quad (6.1)$$

This design is constructed from orthogonal polynomials of degrees smaller or equal to 4. The j-th column of D represents a polynom of degree $j-1$ (as a function of the row index). A submodel $\mathcal{U} \subseteq \{0, 1, 2, 3, 4\}$ is thus given by a subset of degrees of orthogonal polynomials.

Moreover, we assume

$$\begin{aligned} &\varepsilon(\vartheta_r) = (\varepsilon_1(\vartheta_r), \ldots, \varepsilon_{10}(\vartheta_r), \varepsilon_{11}(\vartheta_r))' \sim \mathcal{N}_{11}(0, I), \\ &\beta(\vartheta_r) \sim \mathcal{N}_5(b, \Sigma_\beta), \ \Sigma_\beta = \operatorname{diag}(\tau_0^2, \ldots, \tau_4^2), \end{aligned} \tag{6.2}$$

where I is the 11×11 identity matrix. The new design point at which prediction is made is

$$c = (1, 1.5, -2.5, -3.5, 0.5)', \tag{6.3}$$

which is 'fairly close' to the 7-th observed design point $(D_{70}, \ldots, D_{74})'$.

The two specifications we consider are

(M1) $n = 10$, $N = 100$, D as in (6.1), c as in (6.3),
and for (6.2): $b = (1, 0, 0, 0, 0)'$, $\Sigma_\beta = \operatorname{diag}(1, 1, 0, 1, 0)$,

(M2) $n = 10$, $N = 5$, D as in (6.1), c as in (6.3),
and for (6.2): $b = (1, 0, 0, 0, 0)'$, $\Sigma_\beta = \operatorname{diag}(1, 0.01, 0, 0.01, 0)$.

Hence, the set of the true regression parameters is in both specifications $\mathcal{T} = \{0, 1, 3\}$. The full model $\{0, 1, 2, 3, 4\}$ is overparametrized but still a good basis model for estimating the structural hyper-parameters and thus for estimating expected losses (relative to this basis model). Some realizations of the model specified by (M1) are given in Figure 6.1.

Because $\mathbb{E}[\beta_j(\vartheta_r)] = 0$ for $j = 1, \ldots, 4$, the individual effects almost disappear in the collective representation in Figure 6.1. As mentioned in section 1, decision making on the collective level can lead to anti-selection in insurance.

For each submodel $\mathcal{U}$ of interest we get approximations for

$$\mathbb{E}\left[\left(\hat{X}_{n+1}^{(\mathcal{U})}(c) - X_{n+1,r}(c)\right)^2\right] \text{ (actual expected loss)},$$

$$\mathbb{E}\left[\hat{L}^{(\mathcal{U})}(c)\right], \ \hat{L}^{(\mathcal{U})}(c) \text{ from (3.4)},$$

$$\mathbb{E}\left[\hat{R}^{(\mathcal{U})}(c)\right], \ \hat{R}^{(\mathcal{U})}(c) \text{ estimated version of (4.5)},$$

by simulating 50 independent realizations (of the whole model specified by (M1) or (M2)). We denote these approximations with the symbol Ave (average) instead of $\mathbb{E}$. Note that there is no functional dependence on the risk r since all volumes are $V_i^{(r)} \equiv 1$.

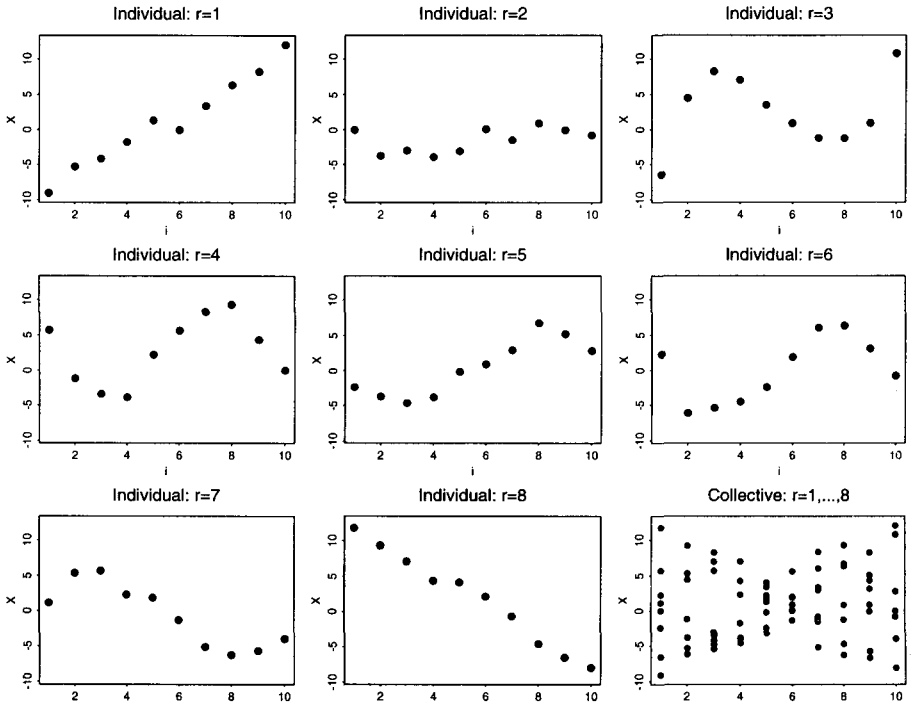

FIGURE 6.1: Eight realizations of individual samples and their joint representation as a collective sample from specification (M1).

For (M1), we considered all $2^5 - 1 = 31$ possible submodels. The results are summarized in Figure 6.2. The left panel of figure 6.2 shows that the actual expected loss $\text{Ave}\left[\left(\hat{X}_{n+1}^{(\mathcal{U})}(c) - X_{n+1,r}(c)\right)^2\right]$ is close to $\text{Ave}\left[\hat{L}^{(\mathcal{U})}(c)\right]$ (the difference between $\text{Ave}\left[\hat{L}^{(\mathcal{U})}(c)\right]$ and $\text{Ave}\left[\hat{R}^{(\mathcal{U})}(c)\right]$ is invisible on this scale and plotted is only the first of these quantities). The four best models are magnified in the right panel of Figure 6.2:

the true model {0, 1, 3},
overparametrized models {0, 1, 2, 3, 4}, {0, 1, 2, 3}, {0, 1, 3, 4}.

$\text{Ave}\left[\hat{R}^{(\mathcal{U})}(c)\right]$ produces the same ranking as the actual expected loss: in particular, minimal $\text{Ave}\left[\hat{R}^{(\mathcal{U})}(c)\right]$ is achieved for the optimal model, being the true one. This is not the case for $\text{Ave}\left[\hat{L}^{(\mathcal{U})}(c)\right]$.

For (M2), the seven most reasonable models are

the true model {0, 1, 3},
overparametrized models {0, 1, 2, 3, 4}, {0, 1, 2, 3}, {0, 1, 3, 4},
underparametrized model {0, 1}, {0, 3}, {0}.

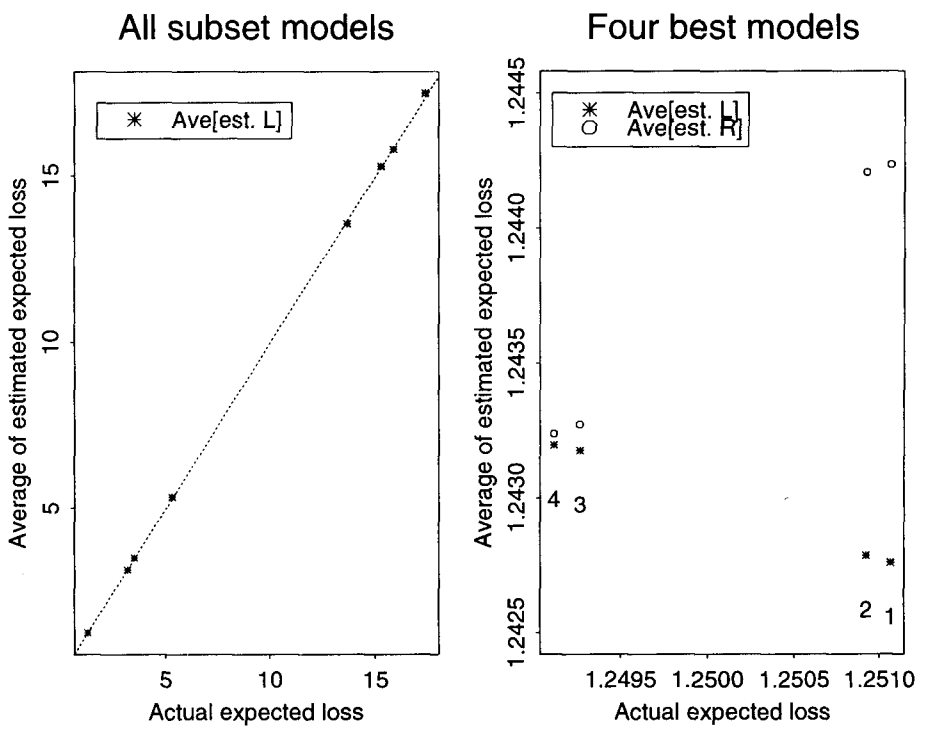

FIGURE 6.2: Left panel: Expected loss and averages of their estimates with (3.4) for all 31 subset models from specification (M1) (every star represents more than one model). The fine dashed reference line is $y = x$. Right panel: Magnification for the best four models. The stars and circles represent averages of estimated expected losses with (3.4) and the plug-in estimate of (4.5), respectively. The models are numbered as: $1 = \{0, 1, 2, 3, 4\}$, $2 = \{0, 1, 2, 3\}$, $3 = \{0, 1, 3, 4\}$, $4 = \{0, 1, 3\}$.

The underparametrized models delete one or both elements in the set of regressor indices $\{1, 3\}$ which corresponds to regression parameters 'close' to zero since $b_1 = b_3 = 0$ and $\tau_1^2 = \tau_3^2 = 0.01$ in (6.2). The results are displayed in Figure 6.3. The optimal model is the pure intercept model $\{0\}$ and not the true model. The estimated expected losses $\text{Ave}\big[\hat{L}^{(\mathcal{U})}(c)\big]$ and $\text{Ave}\big[\hat{R}^{(\mathcal{U})}(c)\big]$ are similar. As in the right panel of Figure 6.2, there is a slight advantage for $\text{Ave}\big[\hat{R}^{(\mathcal{U})}(c)\big]$: it is minimal for the best model and produces the correct ranking, except for model $\{0, 3\}$.

To get an idea about variability and the distribution of the selection rules we consider

$$\delta_L = \hat{L}^{(\mathcal{U}_1)}(c) - \hat{L}^{(\mathcal{U}_2)}(c), \ \delta_R = \hat{R}^{(\mathcal{U}_1)}(c) - \hat{R}^{(\mathcal{U}_2)}(c), \tag{6.4}$$

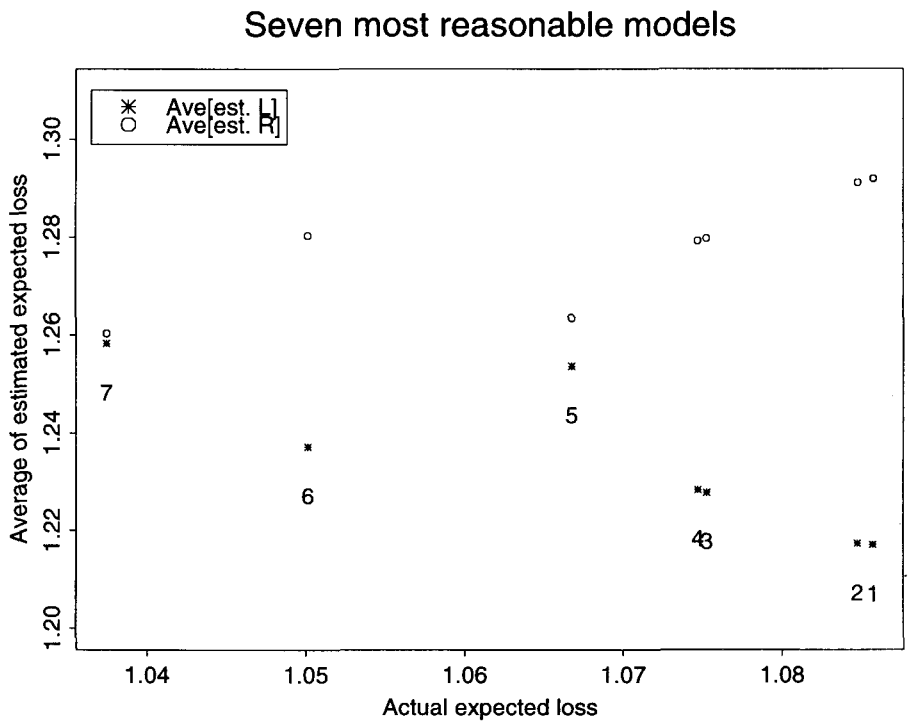

FIGURE 6.3: Expected losses and averages of their estimates for models from specification (M2). The stars and circles represent averages of estimated expected losses with (3.4) and the plug-in estimate of (4.5), respectively. The models are numbered as: 1 = {0, 1, 2, 3, 4}, 2 = {0, 1, 2, 3}, 3 = {0, 1, 3, 4}, 4 = {0, 1, 3}, 5 = {0, 1}, 6 = {0, 3}, 7 = {0}.

for the selection between models $\mathcal{U}_1$ and $\mathcal{U}_2$. Our choices are

$$\mathcal{U}_1 = \{0,\ 1,\ 3\},\ \mathcal{U}_2 = \{0,\ 1,\ 2,\ 3,\ 4\} \text{ for (M1)}, \tag{6.5}$$

$$\mathcal{U}_1 = \{0\},\ \mathcal{U}_2 = \{0,\ 1,\ 2,\ 3,\ 4\} \text{ for (M2)}. \tag{6.6}$$

TABLE 6.1

MISCLASSIFICATION RATES WITH DECISIONS BASED ON δ_L AND δ_R FROM (6.4) FOR THE MODELS IN (6.5) AND (6.6)

	δ_L	δ_R
models from (6.5) for (M1)	0.48	0.12
models from (6.6) for (M2)	0.68	0.30

These are the optimal and full models in both specifications (M1) and (M2). The approximate distributions (estimated from the 50 simulations) of δ_L and δ_R in (6.4) for the models in (6.5) and (6.6) are given in Figure 6.4 in terms of boxplots.

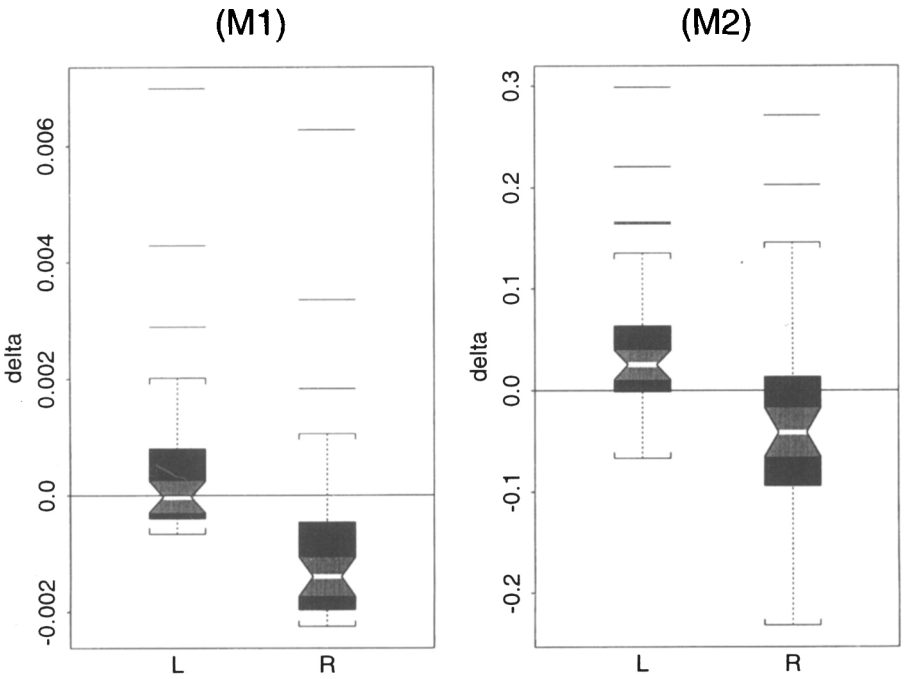

FIGURE 6.4: Boxplot representation for estimated distributions of δ_L and δ_R in (6.4), denoted by L and R, respectively. The two specifications are as in (6.5) and (6.6), denoted by (M1) and (M2), respectively.

Note that a negative value for δ, in (6.4) leads to a correct selection among the two candidate models. Figure 6.4 shows that δ_R is substantially more concentrated on negative values than δ_L (although the variability of δ_R is larger), for both specifications and thus for large and small N. Table 6.1 gives the misclassification rate.

$$\text{MCR} = 50^{-1} \sum_{i=1}^{50} 1_{[\delta_i > 0]},$$

being the relative frequency of misclassifications, where δ_i denotes δ_L or δ_R from (6.4) based on the i-th simulated data-set.

We conclude from Figure 6.4 and Table 6.1 that not only Ave$\left[\hat{R}^{(\mathcal{U})}(c)\right]$ is slightly better than Ave$\left[\hat{L}^{(\mathcal{U})}(c)\right]$ but also the decision rule itself.

The little simulation experiment is reassuring. The averages of the estimated expected losses are almost equal to the true actual losses, even for a small number of risks N. The more complicated estimator $\hat{R}^{(\mathcal{U})}(c)$ has a slightly better behavior in that respect than its simpler cousin $\hat{L}^{(\mathcal{U})}(c)$. The model selector itself based on $\hat{R}^{(\mathcal{U})}(c)$ is able to discriminate reasonably well among very similar prediction models. This is not true for $\hat{L}^{(\mathcal{U})}(c)$ as indicated by Figure 6.4 and Table 6.1. However, model selection based on $\hat{L}^{(\mathcal{U})}(c)$ is accurate if the models are 'sufficiently' different, see also left panel of Figure 6.2. For many practical purposes, it suffices to discriminate among 'sufficiently' different prediction models and selection can then be based on the simpler statistic $\hat{L}^{(\mathcal{U})}(c)$.

7 Discussion

We have developed a framework for model selection in the general credibility regression model (linear model with 'random effects') with collateral data structure. As already known from frequentist statistics, the machinery of hypothesis testing is not tailored towards, and inappropriate for, optimal prediction. The search for optimal prediction models can be done by direct estimation of an expected loss.

Our approach for model selection is to minimize the expected squared error for prediction at an arbitrary (future) design point. We do *not* require specification of a prior distribution for the hyper-parameters. All what we assume is a structure of first and second order moments. In this sense, the approach is 'robust' against misspecification of prior distributions. This issue has always been a main focus in credibility models; it is an analogue to the Gauss-Markov conditions and BLUE estimators in standard linear model theory.

As pointed out in Remark A, the C_p criterion (Mallows, 1973), and others like AIC, BIC or MDL, is not appropriate for the situation encountered here. The reasons are:

- We need to account for variability of (random) parameters.
- We have to study the expected loss for shrinkage estimators.
- We aim for optimal prediction at a (future) design point instead of optimality 'averaged' over the observed design of the data.

The first two issues are major points which need to be considered. The third point is more our preference to do predictive model selection which is optimal at a particular design point. As a result, the optimal model will then depend on the value of this design point. It depends also on the risk r implying that in actuarial applications one might consider the possibility to use different models for different risks.

We have given here three results for the expected loss of the credibility (shrinkage) estimator.

(a) Formula (3.3) describes the exact expected loss for the predictor based on the credibility estimator with known structural parameters, i.e., with known optimal credibility weights.
(b) Formula (4.1) describes the exact expected loss for the predictor based on the credibility estimator with fixed, given credibility weights which are generally not optimal.
(c) Formula (4.5) describes the approximate expected loss for the predictor based on the credibility estimator with estimated credibility weights.

Estimation of these expected losses in (a)-(c) can be done by the plug-in principle. The more complicated nature of the versions in (b) and (c) is the price we pay to get knowledge about more realistic cases than in (a). The version in (b) is also interesting from a theoretical point of view since it gives the *exact, non-asymptotic* expected loss for shrinkage estimation. Note, that the differences between these expected losses are not substantial if N is 'sufficiently' large. Indeed, as $N \to \infty$, all the versions in (a)-(c) are equivalent, and they are exact *regardless* of the size of the individual sample size n. Thus, the most simple, user-friendly criterion in (3.3) often leads to a data-driven model selector in (3.5) which is satisfactory for many practical purposes. To discriminate among very similar prediction models, there can be considerable gain by using $\hat{R}^{(U)}(c)$ instead of $\hat{L}^{(U)}(c)$. Our exploratory simulation study confirms these issues.

The general strategy which we have developed here will also be useful and successful in many other credibility models. For example, the hierarchical models, cf. Jewell (1975) and Taylor (1979), or hierarchical regression models, cf. Sundt (1979) and Norberg (1986).

8 Proofs

Proof of formula (3.1). We make use of the following facts:

(a) $\mathrm{Cov}\left(b_{jr}^X, b_{kr}^X | \vartheta_r\right) = 0$ for $j \neq k$, $r = 1, \ldots, N$.

(b) $\mathrm{Cov}\left(\beta_j(\vartheta_r), \beta_k(\vartheta_r)\right) = 0$ for $j \neq k$, $r = 1, \ldots, N$.

(c) $\mathrm{Cov}(\varepsilon_{n+1}(\vartheta_r), \varepsilon_i(\vartheta_\ell) | \vartheta_r, \vartheta_\ell) = 0$ for $i = 1, \ldots, n$ and $r, \ell = 1, \ldots, N$. This implies uncorrelatedness of $\epsilon_{n+1}(\vartheta_r)$ with the predictor $\hat{X}_{n+1}^{(U)}(c)$.

Note that (a) is due to the orthogonal design which is well known in linear fixed effects regression theory; (b) and (c) are true by assumption. All these issues imply by straightforward calculation formula (3.1).

Proof of formula (3.3). This is just a special case of formula (4.1) when using for $\tilde{Z}_j^{(r)}$ the true optimal weights

$$Z_j^{(r)} = \frac{S_{jj} V_\bullet^{(r)}}{S_{jj} V_\bullet^{(r)} + \sigma^2/\tau_j^2} .$$

Straightforward calculation then yields (3.3). □

Proof of formula (4.1). We take formula (3.2) as our starting point. We first analyse $\mathbb{E}\left[\left(\beta_j(\vartheta_r) - \hat{\beta}_j(\vartheta_r)\right)^2\right]$ appearing in the second term on the right hand side of (3.2). The calculation is straightforward, using again issues (a)-(c) from the proof of formula (3.1) above. The steps are:

$$\begin{aligned}
\mathbb{E}\left[\left(\beta_j(\vartheta_r) - \hat{\beta}_j(\vartheta_r)\right)^2\right] &= \mathbb{E}\left[\left(\tilde{Z}_j^{(r)}\left(\beta_j(\vartheta_r) - b_{jr}^X\right) + (1-\tilde{Z}_j^{(r)})\left(\beta_j(\vartheta_r) - \frac{\sum_{\ell=1}^N b_{j\ell}^X \tilde{Z}_j^{(\ell)}}{\tilde{Z}_j^{(\bullet)}}\right)\right)^2\right] \\
&= \mathbb{E}\left[\left(\left(\tilde{Z}_j^{(r)} + \frac{1-\tilde{Z}_j^{(r)}}{\tilde{Z}_j^{(\bullet)}}\tilde{Z}_j^{(r)}\right)\left(\beta_j(\vartheta_r) - b_{jr}^X\right) + (1-\tilde{Z}_j^{(r)})\sum_{\ell\neq r}\frac{\tilde{Z}_j^{(\ell)}}{\tilde{Z}_j^{(\bullet)}}\left(\beta_j(\vartheta_r) - b_{j\ell}^X\right)\right)^2\right] \\
&= \left(\tilde{Z}_j^{(r)} + \frac{1-\tilde{Z}_j^{(r)}}{\tilde{Z}_j^{(\bullet)}}\tilde{Z}_j^{(r)}\right)^2 \sigma_{j,r}^2 + (1-\tilde{Z}_j^{(r)})^2 \mathbb{E}\left[\left(\sum_{\ell\neq r}\frac{\tilde{Z}_j^{(\ell)}}{\tilde{Z}_j^{(\bullet)}}\left(\beta_j(\vartheta_r) - b_{j\ell}^X\right)\right)^2\right] \\
&= I + (1-\tilde{Z}_j^{(r)})^2 II,
\end{aligned} \tag{8.1}$$

with $\sigma_{j,r}^2$ as defined in (4.1). The first term I is already as it appears in (4.1). For the second term we obtain

$$\begin{aligned}
II &= \mathbb{E}\left[\sum_{\ell,m\neq r}\frac{\tilde{Z}_j^{(\ell)}\tilde{Z}_j^{(m)}}{\left(\tilde{Z}_j^{(\bullet)}\right)^2}\left(\beta_j(\vartheta_r) - b_{j\ell}^X\right)\left(\beta_j(\vartheta_r) - b_{m\ell}^X\right)\right] \\
&= \sum_{\ell\neq r}\frac{\left(\tilde{Z}_j^{(\ell)}\right)^2}{\left(\tilde{Z}_j^{(\bullet)}\right)^2}\mathbb{E}\left[\left(\beta_j(\vartheta_r) - b_{j\ell}^X\right)^2\right] \\
&\quad + \sum_{\substack{\ell\neq r, m\neq r \\ l\neq m}}\frac{\tilde{Z}_j^{(\ell)}\tilde{Z}_j^{(m)}}{\left(\tilde{Z}_j^{(\bullet)}\right)^2}\mathbb{E}\left[\left(\beta_j(\vartheta_r) - b_{j\ell}^X\right)\left(\beta_j(\vartheta_r) - b_{m\ell}^X\right)\right] .
\end{aligned}$$

Now use that

$$\begin{aligned}
&\mathbb{E}\left[\left(\beta_j(\vartheta_r) - b_{j\ell}^X\right)^2\right] = \sigma_{j,\ell}^2 + 2\tau_j^2 \text{ for } \ell \neq r, \\
&\mathbb{E}\left[\left(\beta_j(\vartheta_r) - b_{j\ell}^X\right)\left(\beta_j(\vartheta_r) - b_{jm}^X\right)\right] = \tau_j^2 \text{ for } \ell \neq r, \ \ m \neq r, \text{ and } l \neq m .
\end{aligned}$$

Then,

$$II = \frac{\sum_{\ell \neq r} \left(\tilde{Z}_j^{(\ell)}\right)^2 \left(\sigma_{j,\ell}^2 + \tau_j^2\right) + \left(\sum_{\ell \neq r} \tilde{Z}_j^{(\ell)}\right)^2 \tau_j^2}{\left(\tilde{Z}_j^{(\bullet)}\right)^2}.$$

This, together with (8.1) gives the formula for $\mathbb{E}\left[\left(\beta_j(\vartheta_r) - \hat{\beta}_j(\vartheta_r)\right)^2\right]$ and hence for the second term on the right hand side of formula (4.1).

For the third term on the right hand side of (3.2), it is easy to see that due to issue (b) in the proof of formula (3.1) above,

$$\mathbb{E}\left[\left(\sum_{j \in \mathcal{U}^C} c_j \beta_j(\vartheta_r)\right)^2\right] = \left(\sum_{j \in \mathcal{U}^C} c_j b_j\right)^2 + \sum_{j \in \mathcal{U}^C} c_j^2 \tau_j^2.$$

This then completes the proof of formula (4.1). □

Derivation of formula (4.4). Write $Z_j = \frac{S_{jj} nV}{S_{jj} nV + \sigma^2/\tau_j^2} = g(\sigma^2, \tau_j^2)$. A first order Taylor expansion of g at (σ^2, τ_j^2) yields,

$$\tilde{Z}_j \approx Z_j + \frac{\partial g}{\partial \sigma^2}(\sigma^2, \tau_j^2)(\hat{\sigma}^2 - \sigma^2) + \frac{\partial g}{\partial \tau_j^2}(\sigma^2, \tau_j^2)(\hat{\tau}_j^2 - \tau_j^2). \tag{8.2}$$

The partial derivatives are

$$\frac{\partial g}{\partial \sigma^2}(\sigma^2, \tau_j^2) = -Z_j(1 - Z_j)/\sigma^2, \quad \frac{\partial g}{\partial \tau_j^2}(\sigma^2, \tau_j^2) = Z_j(1 - Z_j)/\tau_j^2. \tag{8.3}$$

All we need to do is then to calculate the variances and covariances of $\hat{\sigma}^2$ and $\hat{\tau}_j^2$. Assuming $\varepsilon_i(\vartheta_r)$ independent of ϑ_r and normality of β_j $(j = 0, \ldots, p-1)$ and of ε_i (normality is only crucial for the values of fourth moments), the following result can be derived from the ideas in Klotz et al. (1969),

$$\hat{\sigma}^2 = \frac{\sigma^2}{N(n-p)} U_1, \; U_1 \sim \chi^2_{N(n-p)},$$

$$\hat{\tau}_j^2 = \left(\frac{\sigma^2}{S_{jj} nV} + \tau_j^2\right) \frac{1}{N-1} U_2 - \frac{\hat{\sigma}^2}{S_{jj} nV}, \; U_2 \sim \chi^2_{N-1},$$

where U_1 and U_2 are independent. Hence,

$$\begin{aligned}
\mathrm{Var}(\hat{\sigma}^2) &= \frac{2\sigma^4}{N(n-p)}, \\
\mathrm{Var}(\hat{\tau}_j^2) &= \frac{2\sigma^4}{(S_{jj}nV)^2 N(n-p)} + \frac{2}{N-1}\left(\frac{\sigma^2}{S_{jj}nV} + \tau_j^2\right)^2, \qquad (8.4) \\
\mathrm{Cov}(\hat{\sigma}^2, \hat{\tau}_j^2) &= -\frac{2\sigma^4}{S_{jj}nV\,N(n-p)}.
\end{aligned}$$

By (8.2),

$$\begin{aligned}
\mathrm{Var}(\hat{Z}_j) \approx{}& \left(\frac{\partial g}{\partial \sigma^2}(\sigma^2, \tau_j^2)\right)^2 \mathrm{Var}(\hat{\sigma}^2) + \left(\frac{\partial g}{\partial \tau_j^2}(\sigma^2, \tau_j^2)\right)^2 \mathrm{Var}(\hat{\tau}_j^2) \\
&+ 2\left(\frac{\partial g}{\partial \sigma^2}(\sigma^2, \tau_j^2)\right)\left(\frac{\partial g}{\partial \tau_j^2}(\sigma^2, \tau_j^2)\right)\mathrm{Cov}(\hat{\sigma}^2, \hat{\tau}_j^2).
\end{aligned}$$

Inserting (8.3) and (8.4) yields (4.4): for the case with $\tau_j^2 = 0$ we take the limit as $\tau_j^2 \to 0$. □

Unbiasedness of $\hat{\sigma}^2$ in (5.1). Denote by

$$\eta_j(\vartheta_r) = b_{jr}^X - \beta_j(\vartheta_r) = \frac{\sum_{i=1}^n \varepsilon_i(\vartheta_r) D_{ij} V_i^{(r)}}{S_{jj} V_\bullet^{(r)}}, \; j = 0, \ldots, p-1; \; r = 1, \ldots, N.$$

Then, by the assumptions on the $\varepsilon_i(\vartheta_r)$'s

$$\mathbb{E}\left[\eta_j(\vartheta_r)|\vartheta_r\right] = 0, \; \mathbb{E}\left[\eta_j^2(\vartheta_r)|\vartheta_r\right] = \frac{\sigma^2(\vartheta_r)}{S_{jj} V_\bullet^{(r)}}.$$

Therefore, the residual sum of squares as defined preceding (5.1) can be written as

$$\begin{aligned}
\mathrm{RSS}_r^2 &= \sum_{i=1}^n \left(\varepsilon_i(\vartheta_r) - \sum_{j=0}^{p-1} \eta_j(\vartheta_r) D_{ij}\right)^2 \frac{V_i^{(r)}}{V_\bullet^{(r)}} \\
&= \sum_{i=1}^n \varepsilon_i^2(\vartheta_r) \frac{V_i^{(r)}}{V_\bullet^{(r)}} + \sum_{j=0}^{p-1} \sum_{i=1}^n \eta_j^2(\vartheta_r) D_{ij}^2 \frac{V_i^{(r)}}{V_\bullet^{(r)}} - 2 \sum_{j=0}^{p-1} \sum_{i=1}^n \eta_j(\vartheta_r) \varepsilon_i(\vartheta_r) D_{ij} \frac{V_i^{(r)}}{V_\bullet^{(r)}} \\
&= I + II + III.
\end{aligned}$$

By definition of S_{jj} and $\eta_j(\vartheta_r)$,

$$II = \sum_{j=0}^{p-1} \eta_j^2(\vartheta_r) S_{jj}, \; III = -2 \sum_{j=0}^{p-1} \eta_j^2(\vartheta_r) S_{jj}.$$

Hence,

$$\begin{aligned} \mathbb{E}\left[\mathrm{RSS}_r^2 | \vartheta_r\right] &= \mathbb{E}\left[\sum_{i=1}^{n} \varepsilon_i^2(\vartheta_r) \frac{V_i^{(r)}}{V_\bullet^{(r)}} - \sum_{j=0}^{p-1} \eta_j^2(\vartheta_r) S_{jj} \Big| \vartheta_r \right] \\ &= n\sigma^2(\vartheta_r) \frac{1}{V_\bullet^{(r)}} - \sum_{j=0}^{p-1} \sigma^2(\vartheta_r) \frac{1}{V_\bullet^{(r)}} = \frac{\sigma^2(\vartheta_r)(n-p)}{V_\bullet^{(r)}}. \end{aligned}$$

But this implies

$$\mathbb{E}\left[\mathrm{RSS}^2\right] = \mathbb{E}\left[\sum_{r=1}^{N} \frac{\sigma^2(\vartheta_r)(n-p) V_\bullet^{(r)}}{V_\bullet^{(r)} V_\bullet^{(\bullet)}} \right] = \frac{\sigma^2(n-p)N}{V_\bullet^{(\bullet)}},$$

which proves unbiasedness of $\hat{\sigma}^2$.

Derivation of the estimator in (5.3). We calculate the expected values $\mathbb{E}\left[W_j\right]$, $j = 0, \ldots, p-1$. Without loss of generality we assume $b_j = E\left[\beta_j(\vartheta_r)\right] = 0$. Straightforward calculation yields,

$$\mathbb{E}\left[W_j | \vartheta_1, \ldots, \vartheta_N\right] = \sum_{r=1}^{N} \left(\beta_j^2(\vartheta_r) + \frac{\sigma^2(\vartheta_r)}{S_{jj} V_\bullet^{(r)}} \right) \frac{V_\bullet^{(r)}}{V_\bullet^{(\bullet)}} - \left(\sum_{\ell=1}^{N} \beta_j(\vartheta_\ell) \frac{V_\bullet^{(\ell)}}{V_\bullet^{(\bullet)}} \right)^2 - \sum_{\ell=1}^{N} \frac{\sigma^2(\vartheta_\ell)}{S_{jj} V_\bullet^{(\ell)}} \left(\frac{V_\bullet^{(\ell)}}{V_\bullet^{(\bullet)}} \right)^2.$$

Thus,

$$\begin{aligned} \mathbb{E}\left[W_j\right] &= \tau_j^2 + N \frac{\sigma^2}{S_{jj} V_\bullet^{(\bullet)}} - \sum_{\ell=1}^{N} \left(\frac{V_\bullet^{(\ell)}}{V_\bullet^{(\bullet)}} \right)^2 \tau_j^2 - \frac{\sigma^2}{S_{jj} V_\bullet^{(\bullet)}} \\ &= \tau_j^2 A + (N-1) \frac{\sigma^2}{S_{jj} V_\bullet^{(\bullet)}}, \end{aligned}$$

where $A = 1 - \sum_{\ell=1}^{N} \left(\frac{V_\bullet^{(\ell)}}{V_\bullet^{(\bullet)}} \right)^2$. This then leads to the estimator in (5.3). □

ACKNOWLEDGMENTS

We thank Alois Gisler for constructive discussions as well as a referee who has motivated us to improve an earlier version of the paper.

REFERENCES

[1] AKAIKE, H. (1970). Statistical predictor identification. Annals of the Institute of Statistical Mathematics 22, 203-217.
[2] AKAIKE, H. (1973). Information theory and the maximum likelihood principle. In 2nd International Symposium on Information Theory (Eds. B.N. Petrov & F. Csàki), Akademiai Kiàdo, Budapest.
[3] BÜHLMANN, H. and GISLER, A. (1997). Credibility in the regression case revisited (A late tribute to Charles A. Hachemeister). ASTIN Bulletin 27, 83-98.
[4] GELFAND, A. and GOSH, S. (1998). A minimum posterior predictive loss approach. Biometrika 85, 1-11.
[5] HACHEMEISTER, C.A. (1975). Credibility for regression models with application to trend. In Credibility, theory and application, Proc. of the Berkeley Actuarial Research Conference on Credibility, pp. 129-163. Academic Press.
[6] JEWELL, W.S. (1975). The use of collateral data in credibility theory: a hierarchical model. Giornale dell'Instituto Italiano degli Attuari 38, 1-16.
[7] KLOTZ, J.H., MILTON, R.C. and ZACKS, S. (1969). Mean square efficiency of estimators of variance components. Journal of the American Statistical Association, 64, 1383-1402.
[8] LINHART, H. and ZUCCHINI, W. (1986). Model Selection. Wiley.
[9] MALLOWS, C.L. (1973). Some comments on C_p. Technometrics 15, 661-675.
[10] NEUHAUS, W. (1985). Choice of statistics in linear Bayes estimation. Scandinavian Actuarial Journal, 1-26.
[11] NORBERG, R. (1986). Hierarchical credibility: analysis of a random effect linear model with nested classification. Scandinavian Actuarial Journal, 204-222.
[12] RISSANEN, J. (1989). Stochastic Complexity in Statistical Inquiry. World Scientific.
[13] SCHWARZ, G. (1978). Estimating the dimension of a model. Annals of Statistics 6, 461-464.
[14] SUNDT, B. (1979). A hierarchical regression credibility model. Scandinavian Actuarial Journal, 107-114.
[15] TAYLOR, G.C. (1979). Credibility analysis of a general hierarchical model. Scandinavian Actuarial Journal, 1-12.
[16] WEISBERG, S. (1985). Applied Linear Regression. Wiley.

Seminar für Statistik
ETH Zürich
CH-8092 Zürich
Switzerland
E-mail: buhlmann@stat.math.ethz.ch

Departement Mathematik
ETH Zürich
CH-8092 Zürich
Switzerland
E-mail: hbuhl@math.ethz.ch